ROYAL AIRCRAFT ESTABLISHMENT

*

Technical Report 81081

July 1981

AN EXAMINATION OF STABILITY CRITERIA FOR ITERATIVE NUMERICAL SCHEMES USED IN SOLVING DIFFERENTIAL EQUATIONS

by

Katharine Moore

*

Procurement Executive, Ministry of Defence
Farnborough, Hants

UDC 517.956.224 : 517.962.2 : 519.6 : 518.1 : 517.933

ROYAL AIRCRAFT ESTABLISHMENT

Technical Report 81081

Received for printing 6 July 1981

AN EXAMINATION OF STABILITY CRITERIA FOR ITERATIVE NUMERICAL SCHEMES USED IN SOLVING DIFFERENTIAL EQUATIONS

by

Katharine Moore

SUMMARY

The stability of numerical schemes for solving algebraic finite-difference equations resulting from finite-difference approximations to differential equations is discussed. It is suggested that the von Neumann method together with its stability criterion provides a reasonably simple way of determining stability. However, there are limitations in its applicability, some of which are indicated. The method is tested in two examples and an indication is given of how best to treat first- and mixed-derivative terms occurring in differential equations.

Departmental Reference: Aero 3508

LIST OF CONTENTS

1 INTRODUCTION

The mathematical modelling of practical problems often involves the use of differential equations. Very few of these equations can be solved analytically, and hence it is of great importance to develop satisfactory schemes for solving them numerically. One requirement of any satisfactory numerical scheme is that it should be stable. However there are several definitions of stability in the literature, all leading to different stability criteria. The purpose of this Report is to discuss some of these definitions and criteria, with the aid of examples. The examples will also suggest how best to treat first-and mixed-derivative terms when they arise in differential equations.

There are two basic steps in the usual finite-difference methods of obtaining approximate numerical solutions to differential equations. First a finite-difference approximation to the differential equation must be chosen, the result of this being a set of equations, termed the finite-difference equations. The second step is to solve these equations, thus obtaining an approximation to the solution of the original differential equation.

Associated with these two steps are four concepts: consistency, convergence, convergence of an iterative scheme, and stability. These will be explained:

Consistency If the finite-difference formulation is equal to the original differential equation plus terms which tend to zero as the grid size tends to zero, then the finite-difference approximation is consistent.

Convergence When the finite-difference approximation is convergent, the difference between the discrete approximation (the solution of the finite-difference equations) and the true solution (the solution of the differential equation) can be made as small as desired, by choosing a sufficiently small grid.

Convergence of an iterative scheme If the finite-difference equations are solved by an iterative scheme, this scheme is convergent if and only if the sequence of approximations (the $(n + 1)$th of which is obtained from the nth by the iterative scheme) converges to the solution of the finite-difference equations. The zeroth approximation, or initial guess, must be supplied.

Stability There seems to be much disagreement over the definition of this term. Some authors merely require that all errors should eventually be damped out[1]. Others appear to relate stability to the growth of rounding errors[2,3].

For initial-value problems Richtmyer and Morton[1] give several definitions of stability. These do not mention errors, as they are given in terms of the operators by which the solution at time $t + \Delta t$ may be obtained from the solution at time t, where 'time' is the coordinate in the marching direction, and Δt the time step taken. These 'operator' definitions can be extended to iterative schemes for solving sets of equations arising from finite-difference approximations to elliptic partial differential equations, if the time-dependent analogy of Jameson (see for instance, Ref 4) is used. In this analogy an artificial time coordinate, t, with step length Δt, in introduced, and ϕ^n, the approximation to the solution after the nth step, is regarded as the solution at time $t = n\Delta t$.

The last two ideas depend only on the finite-difference equations and the manner in which they are solved, no reference being made to the original differential equation. They ensure that the solution of the finite-difference equations can be obtained.

There are many theorems which show that convergence (and convergence of the iterative scheme if appropriate) will be obtained for various classes of differential equations with fairly general initial and boundary conditions, if the finite-difference equations are consistent and the scheme adopted for their solution is stable in a suitable sense (see, for instance, Ref 1). However, for nonlinear second-order partial differential equations there are no such results except in a few special cases. Nevertheless for such equations it is widely accepted (and will be assumed here from now on) that stability (in some suitable sense) and consistency do imply convergence (and convergence of the iterative scheme if appropriate).

It is usually straightforward to show consistency - in practice this is done in the formulation of the finite-difference approximation. In section 2 the discussion on stability will be continued and it is suggested that the von Neumann criterion is a reasonably simple one which is adequate in many cases. Attention is also drawn to some of its limitations. In section 3 some of the ideas introduced in section 2 will be illustrated with specific examples, and it is shown how first- and mixed-derivative terms might best be treated.

2 STABILITY OF ITERATIVE SCHEMES

For many types of differential equations (for example, when the coefficients of the highest derivative terms are functions of the solution) there are no rigorous theories concerning the stability of numerical schemes for their solution. The usual approach is to do a local stability analysis, and to hope that if the scheme is everywhere locally stable it will be globally stable. Although this is not always true, practical experience suggests the correspondence is close, probably because instabilities usually arise locally[1].

There are two methods commonly used for examining the notion of stability of a finite-difference scheme[3]. The first one, termed the von Neumann method, will be examined in section 2.1. The second one, termed the matrix method, will be discussed in section 2.4

2.1 The von Neumann stability method

The differential equation is taken to have constant coefficients, and the problem is assumed to be an initial value one, the only permitted boundary conditions being ones which can be replaced by periodicity conditions. The 'amplification factors', λ, can then be determined as follows:

(i) Substitute, into the finite-difference equations used to solve the differential equation,

$$\underset{\sim}{u}^n(\underset{\sim}{x}) = \lambda^n \underset{\sim}{u}_0 \exp(i\underset{\sim}{k} . \underset{\sim}{x})$$

where $\underline{k}$ is a real vector chosen to satisfy the periodicity conditions, and $\underset{\sim}{u}^n(\underline{x})$ is the vector of unknowns at position $\underline{x}$ after the nth step.

(ii) Determine the amplification factors from the condition for the existence of a non-zero solution for $\underset{\sim}{u}_0$.

von Neumann's criterion states that a necessary condition for stability (as defined in Richtmyer and Morton[1], Chapter 3) is that all the amplification factors, λ , must satisfy

$$|\lambda| \leq 1 + 0(h) \qquad \text{for } 0 < h < \tau$$

where τ is some upper bound on h , and h is the step length in the marching direction. This stability criterion can be extended to the iterative methods for solving elliptic partial differential equations with periodic boundary conditions, by use of Jameson's time-dependent analogy (see, for instance, Ref 4). It then becomes

$$|\lambda| \leq 1 \ .$$

The application of the von Neumann criterion will be illustrated by the following example[1], in which the boundary conditions are periodic:

$$\text{solve } \frac{\partial u}{\partial t} = \sigma \frac{\partial^2 u}{\partial x^2} \quad \text{on } 0 \leq x \leq 1 \quad \text{and} \quad 0 \leq t$$

with σ a constant, $u(0,t) = u(1,t) = 0$ and $u(x,0) = F(x)$. Let u_j^n be the finite-difference solution at $x = j\Delta x$ and $t = n\Delta t$, take a central-difference approximation for $\partial^2 u/\partial x^2$, and use an explicit scheme (that is one in which one unknown u_j^{n+1} is expressed in terms of the known u_j^ns). The finite-difference equations may then be written

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = \frac{\sigma}{(\Delta x)^2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right)$$

giving the equations, for the amplification factors, λ ,

$$\frac{(\lambda - 1)}{\Delta t} = \frac{\sigma}{(\Delta x)^2}\left[e^{ik\Delta x} + e^{-ik\Delta x} - 2\right]$$

where $k = 2\pi n$, with n any integer, to satisfy the periodicity condition. Hence

$$\lambda = 1 - \frac{\sigma\Delta t}{(\Delta x)^2}\, 2[1 - \cos k\Delta x]$$

and the von Neumann criterion gives

$$\frac{2\sigma\Delta t}{(\Delta x)^2} \leq 1 \qquad (2\text{-}1)$$

as a necessary condition for stability.

Under some circumstances the von Neumann condition is sufficient as well as necessary for stability. In particular, for two-level schemes (that is schemes in which the finite-difference equations relate values of the dependent variables at the $(n + 1)$th step to values at the nth step, values of the dependent variables at earlier steps not occurring in the equations) with one dependent variable and any number of independent variables, the von Neumann criterion is sufficient as well as necessary for stability. For the example given above equation (2-1) is thus a necessary and sufficient condition to ensure stability.

In cases where the von Neumann criterion is necessary but not sufficient for stability, further criteria which ensure stability can often be found[1].

Although it might appear here that the von Neumann method has limited applicability, particularly because of the restriction to periodic boundary conditions, in fact it has much wider practical application. In section 2.4.2 the Godunov-Ryabenkii criterion is described. This in effect provides an extension to the von Neumann treatment to cover consideration of arbitrary boundary conditions, although in practice the extension is often difficult to apply.

2.2 Matrix formulation of the finite-difference problem

The problem of finding the solution, ϕ, of a set of finite-difference equations approximating a differential equation (ordinary or partial) can be expressed in the form:

$$\text{find } \phi \text{ satisfying } A\phi = B \tag{2-2}$$

where B is independent of ϕ, but A may depend on ϕ. As an example consider the numerical solution of

$$a\Phi_{xx} + 2b\Phi_{xy} + C\Phi_{yy} = f$$

on a unit square, with $\Delta x = \Delta y = 1/N$, a, b, c, f constants and values of Φ given on the boundary. Let the subscripts, i and j, refer to the coordinate directions, x and y, respectively. Take the usual central-difference representations of Φ_{xx} and Φ_{yy} and assume that if Φ were known, Φ_{xy} would be approximated by

$$\Phi_{xy}(i\Delta x, j\Delta y) = \frac{1}{4\Delta x\Delta y}\left(\Phi_{i+1,j+1} - \Phi_{i+1,j-1} - \Phi_{i-1,j+1} + \Phi_{i-1,j-1}\right) \tag{2-3}$$

where $\Phi_{ij} = \Phi(i\Delta x, j\Delta y)$.

The equation (2-2) takes the form

$$\begin{pmatrix} E & F & & \mathbf{0} \\ F^T & E & \ddots & \\ & \ddots & \ddots & F \\ \mathbf{0} & & F^T & E \end{pmatrix} \begin{pmatrix} \Psi_1 \\ \vdots \\ \Psi_{N-1} \end{pmatrix} = B \tag{2-4}$$

where each Ψ_j is a vector of length $(N-1)$. The ith component of the vector Ψ_j is $\phi\left[i + (j-1)(N-1)\right]$ and is the estimate which is obtained for ϕ_{ij} from the finite-difference approximation. B is a vector of length $(N-1)^2$ with elements equal to f/N^2 plus (possibly) some contribution from the boundary. E and F are matrices of order $(N-1) \times (N-1)$ with

$$E = \begin{pmatrix} -2a-2c & a & & & \\ a & -2a-2c & a & & \\ & \ddots & \ddots & \ddots & \\ & & a & -2a-2c & a \\ & & & a & -2a-2c \end{pmatrix}$$

and

$$F = \begin{pmatrix} c & b/2 & & & \\ -b/2 & c & b/2 & & \\ & \ddots & \ddots & \ddots & \\ & & -b/2 & c & b/2 \\ & & & -b/2 & c \end{pmatrix} \qquad (2\text{-}5)$$

F^T denotes the transpose of F, that is $F^T_{ij} = F_{ji}$. The particular form of the matrix, A, (which is of order $(N-1)^2 \times (N-1)^2$) is determined by the finite-difference approximations used.

The matrix, A, has several features of interest. It is sparse, which means that most of its entries are zero - there is no precise definition of sparse, but a good guide seems to be that a matrix is sparse if more than 90% of its entries are zero. It is block tri-diagonal, and each non-zero block is tri-diagonal. Features of this nature are generally observed with finite-difference equations obtained from partial differential equations.

2.3 Methods of solution

Consider equation (2-2) again. For any given A this set of equations can be solved directly by Gaussian elimination. Frequently some form of reordering of A before carrying out the elimination will give a much faster scheme. This sort of scheme will be most suitable when A is independent of ϕ (as will occur if the differential equation is linear), or if A is triangular (either upper or lower) with all elements on the leading diagonal independent of ϕ (as might occur if the differential equation is hyperbolic or parabolic).

However, the equations can also be solved iteratively by writing A in the form

$$A = G - C$$

and solving

$$Gw^{n+1} = B + Cw^{n} \tag{2-6}$$

where w^0 is given, G is some easily invertible matrix and any ϕ_i appearing in G and C is replaced by the estimate w_i^n of ϕ_i. If the scheme is stable in the sense of all errors decaying suitably then w^n will tend to ϕ as $n \to \infty$. An iterative scheme of this sort will obviously be most suitable when A is dependent on ϕ, and equation (2-2) cannot be solved directly, as might occur if the differential equation is elliptic and 'quasi-linear'. A quasi-linear differential equation is one in which the highest derivatives appear linearly, but the coefficients of the highest derivatives are functions of the dependent variable and lower-order derivatives of it.

It is of interest to note that approximating some second-order parabolic and first-order hyperbolic quasi-linear partial differential equations by certain 'marching' finite-difference schemes can also give rise to equations of the form (2-6), where w^n is now the approximation to the solution at the nth step in the marching direction. This is illustrated in the example described in subsection 2.4.2. Equation (2-2) now takes the form

$$\begin{pmatrix} -C & G & & & \\ & -C & \ddots & & \\ & & \ddots & \ddots & \\ & & & -C & G \end{pmatrix} \begin{pmatrix} w^0 \\ w^1 \\ \vdots \\ w^n \\ \vdots \end{pmatrix} = \begin{pmatrix} B \\ B \\ \vdots \\ B \\ \vdots \end{pmatrix}$$

2.4 Matrix method of determining stability

2.4.1 Preliminaries

It was indicated at the beginning of this section, that there are very few theories concerning stability, when the matrix, A, of equation (2-2) depends on the solution of the finite-difference equations. A local stability analysis is usually carried out in such cases. Hence, from now on, the matrices A of equation (2-2), and G and C of equation (2-6) will be assumed to be independent of the solution, ϕ, of the finite difference equations.

If the 'error', e^n, is now defined by

$$e^n = \phi - w^n$$

it can easily be seen that

$$Ge^{n+1} = Ce^{n} \tag{2-7}$$

and that w^n will tend to ϕ if and only if e^n tends to zero, as n tends to infinity. Thus, it might be expected that, in principle at least, stability could be discussed in terms of the properties of the matrix $G^{-1}C$.

Before discussing some possible stability criteria it is necessary to define a few terms and some notation:

$|\underline{x}| = \sqrt{\sum_{i=1}^{n} x_i x_i^*}$, where $\underline{x}$ is a vector of length n , and $*$ denotes complex conjugate.

Hermitian A square matrix, A , is hermitian if $A_{ji} = A_{ij}^*$.

Diagonally dominant An $n \times n$ complex matrix, A , is diagonally dominant if

$$|A_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |A_{ij}| \qquad \text{for all } i .$$

Negative definite An hermitian $n \times n$ matrix, A , is negative definite if, for all vectors $\underline{x}$ not identically zero, and of length n

$$\sum_{i=1}^{n} \sum_{i=1}^{n} x_i^* A_{ij} x_j < 0 .$$

Eigenvalue and corresponding eigenvector λ is said to be an eigenvalue (also termed latent root or characteristic root) of an $n \times n$ matrix A , if there exists a non-zero vector $\underline{e}$ of length n such that

$$A\underline{e} = \lambda \underline{e} .$$

The vector $\underline{e}$, which is determined to at most an arbitrary multiplicative constant, is termed the eigenvector corresponding to the eigenvalue, λ .

Orthogonal Two vectors, $\underline{x}$ and $\underline{y}$, each of length n are orthogonal if

$$\underline{x} \cdot \underline{y} = \sum_{i=1}^{n} x_i y_i = 0 .$$

Spectral radius, $\rho(A)$ The spectral radius of an $n \times n$ complex matrix is $\rho(A)$ where

$$\rho(A) = \max_{1 \leq i \leq n} |\lambda_i|$$

and $\{\lambda_i\}$ is the set of eigenvalues of A .

Spectral norm, $\|A\|$ The spectral norm of an $n \times n$ complex matrix is $\|A\|$

where $\|A\| = \max_{\text{all } \underline{x} \neq \underline{0}} \frac{|Ax|}{|x|}$

and $\underline{x}$ is a vector of length n .

2.4.2 Stability criteria

Three possible criteria will be described in this section and a discussion of them follows in section 2.4.3.

Clearly, if, in equation (2-7) $|e^n| \to 0$ as $n \to \infty$

$$\rho(G^{-1}C) < 1 \ . \tag{2-8}$$

In equation (2-6) this will ensure that errors eventually decay, so stability is achieved in one of the senses of section 1. For the rest of this Report condition (2-8) will be termed the first stability criterion. This criterion seems to be widely used - see, for instance, Refs 5 and 6.

The most obvious shortcoming of this criterion is that for parabolic and hyperbolic equations it will exclude the possibility of solutions which are growing exponentially in the marching direction. To include these, the stability criterion should be relaxed to

$$\rho(G^{-1}C) \leq 1 + O(h)$$

where h is the step length in the marching direction. However, stability, in any of the senses given in section 1, is no longer necessarily achieved.

Examples[1,6,7] show further unsatisfactory features of the stability criterion (2-8). This will be illustrated here with the example[1]:

$$\text{solve } \frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 \qquad \text{on } 0 \leq x \leq 1 \qquad \text{and } 0 \leq t \ ,$$

with a a positive constant, $u(0,t) = 0$ and $u(x,0) = F(x)$. The true solution is $u = F(x - at)$ when $x \geq at$, and zero everywhere else. Take $\Delta x = 1/N$, represent $\partial u/\partial x$ by a backward difference, and use an explicit scheme. This gives

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = \frac{a}{\Delta x}\left(u_j^n - u_{j-1}^n\right)$$

where u_j^n is the calculated value of u at $x = j\Delta x$ and $t = n\Delta t$. Let $r = \Delta t a/\Delta x$, then

$$u_j^{n+1} = (1 - r)u_j^n + ru_{j-1}^n \tag{2-9}$$

giving the matrix G as the identity and C as the $N \times N$ matrix

$$\begin{pmatrix} 1-r & & & & \\ r & 1-r & & & \\ & r & 1-r & & \\ & & \ddots & \ddots & \\ & & & r & 1-r \end{pmatrix}$$

If $r = 3/2$, $\rho(G^{-1}C) = \frac{1}{2}$ and so condition (2-8) is satisfied. However, f an error initially has the form

$$e^0 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ \vdots \end{pmatrix}\varepsilon \text{ , it is found that } e^1 = \begin{pmatrix} -\frac{1}{2} \\ \frac{3}{4} \\ 0 \\ \vdots \\ \vdots \end{pmatrix}\varepsilon, e^2 = \begin{pmatrix} (-\frac{1}{2})^2 \\ 2(\frac{3}{4})(-\frac{1}{2}) \\ (\frac{3}{4})^2 \\ 0 \\ \vdots \end{pmatrix}\varepsilon$$

and in general

$$e_i^q = 0 \qquad i > q + 1$$

$$= (-\tfrac{1}{2})^{q-(i-1)}(\tfrac{3}{4})^{i-1}\binom{q}{i-1}\varepsilon \qquad i \leq q + 1 .$$

For fixed i it can be shown that $|e_i^q|$ has a maximum at $q = 2i - 2$, and that this maximum is an increasing function of i. Hence

$$\max_{i,q} |e_i^q| = \left(\frac{1}{2}\right)^{N-1}\left(\frac{3}{2}\right)^{N-1} \frac{(2N-2)!}{(N-1)!(N-1)!}\,\varepsilon$$

which is asymptotic to, for large N,

$$\frac{3^{N-1}\varepsilon}{(\pi N)^{\frac{1}{2}}}$$

by Stirling's formula. This shows that stability is not achieved in the sense of the effect of a rounding error being bounded as N increases, nor is it achieved in the sense of any of the definitions given in Richtmyer and Morton[1]. In numerical work the scheme will become less and less satisfactory as $\Delta x = 1/N$ becomes smaller and smaller, while r is kept fixed. However, for small enough N, it may appear satisfactory because $\max_{i,q} |ei^q|/\varepsilon$ will not be very large, and errors increasing only slightly before decaying may be acceptable.

Mathematically, the difficulties of errors becoming arbitrarily large before they decay away, as the grid size decreases, seem to arise either through there being fewer essentially distinct eigenvectors than eigenvalues of $G^{-1}C$ (as in the example just discussed), or through the eigenvectors being nowhere near orthogonal[6,7].

To prevent components of any introduced error becoming arbitrarily large as the grid size is decreased, one might, as suggested in Richtmyer and Morton[1], impose

$$\|(G^{-1}C)^{\nu}\| \leq \text{some constant, } \kappa \qquad (2\text{-}10)$$

independent of the integer, ν, and the grid size for all sufficiently small grid sizes. ν is any integer satisfying $0 \leq h\nu \leq H$, where H is some given constant, and h is the step length in the marching direction, which satisfies $0 < h < \tau$, τ being some suitable upper bound. This condition can be extended to iterative methods for solving elliptic partial differential equations, by use of Jameson's time-dependent analogy (see, for instance, Ref 4). It is the same as that given in equation (2-10), except that ν is now any positive integer.

Satisfying this constraint, which will be termed the second stability criterion for the rest of this Report, seems to ensure stability in all the senses given in section 1. However, for this constraint to be satisfactory in practice, the constant, κ, must not be too large, and the grid size must not be too small.

In the example discussed earlier in this section this stability criterion gives $0 \leq r \leq 1$, which is the well-known Courant-Friedrichs-Lewy criterion for the stability of the iterative scheme.

If there is only one space-like independent variable it can be shown that, if the second stability criterion is to be satisfied, it is necessary that the Godunov-Ryabenkii criterion be satisfied. A statement of the Godunov-Ryabenkii criterion and a proof of its necessity can be found in Richtmyer and Morton[1]. It is in two parts, one part being a condition that looks very like a von Neumann criterion. Amplification factors, λ, (local ones if necessary) are found by exactly the same procedure as described in section 2.1. These must satisfy

$$\lim_{h \to 0} |\lambda| \leq 1$$

where h is the step length in the marching direction.

This bound on the sizes of the amplification factors is very similar to, but weaker than, the bound imposed on them in section 2.1. By use of the time-dependent analogy (see, for instance, Ref 4) this becomes

$$|\lambda| \leq 1$$

for iterative methods used in solving elliptic partial differential equations. This bound on the sizes of the amplification factors is the same as that imposed on them in section 2.1.

In the example discussed earlier, equation (2-9) can be used to show that the amplification factors, λ, have the form

$$\lambda = (1 - r) + re^{-ik\Delta x} \quad .$$

Hence, if the Godunov-Ryabenkii criterion is to be satisfied

$$0 \leq r \leq 1 \ .$$

The second part of the Godunov-Ryabenkii criterion is concerned with the boundary conditions, each boundary condition being treated in turn. The implementation of this part of the criterion for the general case is discussed in Richtmyer and Morton[1]. However the discussion can be much simplified if the following restrictions are imposed:

(i) The differential equation has constant coefficients.

(ii) The differential equation has a scalar dependent variable.

(iii) The highest space derivative is at most second order.

(iv) A two-level difference scheme (this term is defined in section 2.1) is adopted.

(v) There is at most one boundary condition at each boundary.

Attention here will thus be confined to this class of problem.

There must be two boundary conditions for a problem in which the highest space derivative is of order two and one for a problem in which the highest space derivative is of order one - it is taken that this is true. For clarity it will be further assumed that the problem has been formulated in terms of the equation satisfied by the errors and the boundary conditions on them. Errors must satisfy homogeneous boundary conditions (*ie* errors do not contribute anything to the boundary conditions), because the Godunov-Ryabenkii criterion is only applicable to problems with linear boundary conditions. If a particular boundary condition on the dependent variable is homogeneous, errors also satisfy this boundary condition.

If, when carrying out a von Neumann stability analysis on the equation describing the behaviour of the errors, $\exp(ik\Delta x)$ is replaced by μ, an equation relating λ and μ results. (To carry out the von Neumann stability analysis this equation must now be solved with $\mu = \exp(ik\Delta x)$, for all possible values of $k\Delta x$.) If the substitutions

$$e^n(j\Delta x) = \lambda^n \mu^j e_0 \qquad (2\text{-}11)$$

where e^n is the error at the nth step after it is introduced and e_0 is a non-zero constant, are now made into the finite-difference equation modelling the boundary condition on the error, at the boundary under consideration, another equation for λ and μ results. These two equations can be solved simultaneously for λ and μ; $|\mu|$ need not necessarily equal unity.

If the value of u, the dependent variable, is specified at the boundary, the error must always be zero at the boundary, and so μ must be zero, because e_0 cannot be zero. This is always stable. In cases when μ is non-zero, e, given in equation (2-11) is only acceptable (physically) if it decays away from the boundary under consideration, into the interior of the space[1]. Hence at a lower boundary only solutions for λ and μ

with $|\mu| < 1$ are acceptable, while at an upper boundary only solutions with $|\mu| > 1$ are acceptable. The boundary condition under consideration is said to be stable if λ, corresponding to an acceptable μ, satisfies

$$\lim_{h \to 0} |\lambda| \leq 1$$

for marching problems, with h the step length in the marching direction, and

$$|\lambda| \leq 1$$

for elliptic problems. The second part of the Godunov-Ryabenkii criterion is satisfied if all boundary conditions are stable.

It is often rather more difficult to ascertain whether this part of the Godunov-Ryabenkii criterion is satisfied, than it is to ascertain whether the 'von Neumann-like' part of the criterion is satisfied. However, in the example earlier in this section, the boundary condition at $x = 0$ cannot lead to any instabilities, because the dependent variable is zero there.

As has already been commented the second stability criterion might not be sufficiently strong. To avoid the possibility that the constant, κ, of equation (2-10) might be too large in practice, one might impose what will, for the rest of this Report, be termed the third stability criterion,

$$\|G^{-1}C\| \leq 1 + O(h) \qquad (2\text{-}12)$$

where h is the step length in the marching direction (which is to be regarded as zero when there is no marching direction). This, however, seems over-restrictive, because it very rarely matters if an error does grow slightly before decaying.

2.4.3 Discussion of criteria

It should be noted that, if the third stability criterion (2-12) is satisfied, then the other two are necessarily satisfied because (2-12) ⇒ (2-10) ⇒ (2-8). On the other hand, if the eigenvectors of $G^{-1}C$ are mutually orthogonal and there are as many eigenvectors as there are rows of $G^{-1}C$, then, if the first stability criterion (2-8) is satisfied, the other two are necessarily satisfied. This is shown by proving

$$(2\text{-}8) \Rightarrow (2\text{-}10) \Rightarrow (2\text{-}12)$$

under such circumstances.

For practical purposes what is desired is a necessary and sufficient condition for stability (in some suitable sense) which is easily applied to any numerical scheme. For most problems it is very difficult to find $\rho(G^{-1}C)$ or $\|(G^{-1}C)^{\nu}\|$ for all permissible integers, ν. For many problems with one space dimension it is also difficult to apply the part of the Godunov-Ryabenkii criterion involving the boundary conditions. (The Godunov-Ryabenkii criterion is only applicable to problems with one space-like independent

variable.) Only an analysis of the form described in section 2.1, leading to a 'von Neumann-like' stability criterion is usually fairly straightforward. It has so far only been indicated that 'von Neumann-like' criteria are applicable under rather restrictive conditions (see sections 2.1 and 2.4.2). However, it is widely accepted[1] that criteria of this form are applicable to many problems not included in the categories described in sections 2.1 and 2.4.2. Intuitively this 'feels right' because, away from the boundaries, such criteria give necessary conditions for local stability, and the correspondence between local stability and global stability is usually close[1]. However, as the Godunov-Ryabenkii criterion indicates, even if, in a von Neumann-like analysis the amplification factors all have moduli less than unity, the scheme may still be unsatisfactory in practice, because of the boundary conditions. This will be illustrated in section 3.2. In contrast, in some cases where the amplification factors do not obey a von Neumann-like criterion the numerical scheme may appear to be satisfactory. This will arise if the grid is sufficiently coarse and $\rho(G^{-1}C) < 1$. An example of this will be given in section 3.1. In these circumstances the scheme will become less and less satisfactory as the grid is refined.

As the only useful criterion that has been suggested (*ie* a von Neumann-like criterion) is a necessary condition for stability, it might be of value to see if useful criteria which are sufficient for stability can be found. As mentioned in section 2.1, conditions which, with the von Neumann criterion, are sufficient for stability for schemes approximating pure initial value problems can be found in Richtmyer and Morton[1]. For problems involving boundary conditions and variable coefficients the most useful method for finding sufficient conditions for stability, and incidentally for indicating suitable numerical schemes, is probably the so-called 'energy method'[1]. However the application of the method usually seems to require much algebra, and often leads to very complicated conditions which are far from necessary.

A suitable practical approach to obtaining stability might thus be to ensure that the von Neumann-like criterion of section 2.4.2 is satisfied locally everywhere.

3 EXAMPLES

Here the ideas of section 2 will be illustrated with two examples. In section 3.1 the representation of a mixed second derivative in a second-order elliptic differential equation will be discussed. In section 3.2 a second-order ordinary differential equation, in which occur terms involving the first derivative, will be considered. Suitable iterative schemes for solving the finite-difference equations will be suggested.

3.1 Example 1 - mixed second derivative

Consider the equation

$$a\Phi_{xx} + 2b\Phi_{xy} + c\Phi_{yy} + d\Phi_{x} = 0$$

on a unit square with Φ zero everywhere on the boundary and $ac > b^2$ (*ie* the equation is elliptic). Without loss of generality take a to be positive. The only solution of the equation is Φ identically zero everywhere. Let the subscripts, i and j, refer to the coordinate directions, x and y, respectively, and let $\Delta x = \Delta y = 1/N$. Take

the usual central difference representations of Φ_{xx}, Φ_{yy} and Φ_x. One possible representation of the cross-derivative was given in equation (2-3). This representation has a truncation error of

$$\frac{1}{6}\left[(\Delta x)^2\Phi_{xxxy} + (\Delta y)^2\Phi_{xyyy}\right] + \text{higher order terms.}$$

Mitchell[3] suggests

$$\Phi_{xy}(i\Delta x, j\Delta y) \cong \frac{1}{2\Delta x\Delta y}\left[\Phi_{i+1,j+1} - \Phi_{i,j+1} - \Phi_{i+1,j} + \Phi_{i,j} + \Phi_{i,j} - \Phi_{i-1,j} - \Phi_{i,j-1} + \Phi_{i-1,j-1}\right] \quad \text{if } b > 0 \quad (3\text{-}1a)$$

with $\Phi_{ij} = \Phi(i\Delta x, j\Delta y)$, which has a truncation error

$$\frac{(\Delta x)^2}{6}\Phi_{xxxy} + \frac{\Delta x\Delta y}{4}\Phi_{xxyy} + \frac{(\Delta y)^2}{6}\Phi_{xyyy} + \text{higher order terms}$$

and

$$\Phi_{xy}(i\Delta x, j\Delta y) \cong \frac{1}{2\Delta x\Delta y}\left[\Phi_{i,j+1} - \Phi_{i-1,j+1} - \Phi_{ij} + \Phi_{i-1,j} + \Phi_{i+1,j} - \Phi_{ij} - \Phi_{i+1,j-1} + \Phi_{i,j-1}\right] \quad \text{if } b < 0 \quad (3\text{-}1b)$$

with a similar truncation error. Mitchell[3] suggests this scheme so that the coefficients of all the unknowns in the resulting finite-difference equations will be positive. He therefore requires $|b| < \min(a,c)$, in the case of zero d. While this sort of condition is necessary for some of the matrix theories of stability that have been developed[3,8], it is *not* usually a necessary condition for stability and is of no use in the situation $b > \min(a,c)$. It is thus worthwhile exploring alternative representations of the cross-derivative.

The third and last representation of the cross-derivative to be considered here will be termed the 'inverse Mitchell' scheme, because it is the same as Mitchell's scheme (given above in equations (3-1)), except that the condition on b is inverted. It is

$$\Phi_{xy}(i\Delta x, j\Delta y) \cong \frac{1}{2\Delta x\Delta y}\left[\Phi_{i,j+1} - \Phi_{i-1,j+1} - \Phi_{i,j} + \Phi_{i-1,j} + \Phi_{i+1,j} - \Phi_{i,j} - \Phi_{i+1,j-1} + \Phi_{i,j-1}\right] \quad \text{if } b > 0 \quad (3\text{-}2a)$$

and

$$\phi_{xy}(i\Delta x, j\Delta y) \cong \frac{1}{2\Delta x \Delta y}\left[\phi_{i+1,j+1} - \phi_{i,j+1} - \phi_{i+1,j} + \phi_{i,j}\right.$$

$$\left. + \phi_{i,j} - \phi_{i-1,j} - \phi_{i,j-1} + \phi_{i-1,j-1}\right] \qquad \text{if } b>0 \quad (3\text{-}2b)$$

These three possible representations for the cross-derivative all lead to finite-difference equations of the form

$$A\phi = 0 \qquad (3\text{-}3)$$

where, as in section 2.2, $\phi[i + (j - 1)(N - 1)]$ is the estimate of Φ at $x = i\Delta x$, $y = j\Delta y$ which is obtained from solving the finite difference equations and A has the form given in equation (2-4). With the cross-derivative given by equation (2-1) E is an $(N - 1) \times (N - 1)$ matrix with

$$E = -\begin{pmatrix} 2a + 2c & -a - d/2N & & & \\ -a + d/2N & \circ & \circ & & \\ & \circ & \circ & \circ & \\ & & \circ & \circ & -a - d/2N \\ & & & -a + d/2N & 2a + 2c \end{pmatrix}$$

and F is given in equation (2-5). With the cross-derivative as given in equations (3-1)

$$E = -\begin{pmatrix} 2a - 2b + 2c & -a + b - d/2N & & & \\ -a + b + d/2N & \circ & \circ & & \\ & \circ & \circ & \circ & \\ & & \circ & \circ & -a + b - d/2N \\ & & & -a + b + d/2N & 2a - 2b + 2c \end{pmatrix}$$

and

$$F = -\begin{pmatrix} -c + b & -b & & & \\ -c + b & \circ & \circ & & \\ & \circ & \circ & \circ & \\ & & \circ & \circ & -b \\ & & & -c + b & -c + b \end{pmatrix}$$

when $b > 0$, and with the cross-derivatives as given in equations (3-2)

$$E = \begin{pmatrix} 2a + 2b + 2c & -a - b - d/2N & \\ -a - b + d/2N & \circ & \\ \circ & & \end{pmatrix}$$

and

$$F = \begin{pmatrix} -c - b & & \\ & \circ & \\ b & & -c - b \\ \circ & & \\ & b & \end{pmatrix}$$

when $b > 0$. The corresponding matrices when $b < 0$ can easily be derived.

Equation (3-3) will be solved iteratively by a successive line over-relaxation (SLOR) scheme, in which all values of ϕ on a line $y =$ constant are updated at the same time. At points before and on the line where ϕ is currently being updated there is a choice as to whether to use the values of ϕ from the current iteration, or values of ϕ found during the previous iteration. The first derivative, Φ_x , will be calculated using values of ϕ found during the previous iteration, while the second derivatives, Φ_{xx} and Φ_{yy} , will be estimated using values of ϕ from the current iteration, whenever possible. For the mixed second derivative, Φ_{xy} , values of ϕ from the current iteration will be used on the line previous to the current one, while on the current line various schemes, depending on the representation of Φ_{xy} , will be investigated. These schemes are as follows:

Scheme I Φ_{xy} is given in equation (2-3). No values of ϕ on the current line are used.

Scheme II Φ_{xy} is given in equations (3-1). The values of ϕ currently being calculated are used everywhere on the current line.

Scheme III Φ_{xy} is given in equations (3-2). The values of φ currently being calculated are used everywhere on the current line.

Scheme IV Φ_{xy} is given in equations (3-1). The value of ϕ currently being calculated is used at the point at which the derivative is centred. Elsewhere on the current line values of ϕ from the previous iteration are used.

Scheme V Φ_{xy} is given in equations (3-2). The value of ϕ currently being calculated is used at the point at which the derivative is centred. Elsewhere on the current line values of ϕ from the previous iteration are used.

Scheme VI ϕ_{xy} is given in equations (3-1). Values of ϕ from the previous iteration are used to approximate the third and fourth terms. Values of ϕ currently being calculated are used in the fifth and sixth terms.

If $d = 0$ and the matrix, E , is determined from equation (2-4) and any one of equations (2-3), (3-1) and (3-2) it is easily seen that A is hermitian and E is negative definite. Then, if any of schemes I, II and III is employed it can be shown that the scheme will be stable, in the sense of all errors eventually decaying, if the relaxation factor, ω , satisfies $0 < \omega < 2$ and the corresponding matrix, A , is negative definite[8]. This, however, seems of little use, as this definition of stability has already been shown to be of limited value (see section 2.4.2). Also d will often not be zero.

A von Neumann type analysis can, however, be used to give some indications of stability in the sense of all the definitions given in section 1. The amplification factors, λ , are as follows, where $\theta = 2\pi k/N$ and $\psi = 2\pi \ell/N$, k and ℓ being integers satisfying $1 \leq k, \ell \leq N - 1$:

Scheme I

$$\lambda = -\frac{2(\omega - 1)a(1 - \cos\theta) + 2c(\omega - 1) - \omega c e^{i\psi} - \frac{d}{N} i\omega \sin\theta - b\omega e^{i\psi} i \sin\theta}{2a(1 - \cos\theta) + 2c - \omega c e^{i\psi} + b\omega e^{i\psi} i \sin\theta}$$

Scheme II $(b > 0)$

$$\lambda = -\frac{2(\omega - 1)(a - b)(1 - \cos\theta) + 2c(\omega - 1) - \omega c e^{i\psi} - \frac{d}{N} i\omega \sin\theta - b\omega e^{i\psi}(e^{i\theta} - 1)}{2(a - b)(1 - \cos\theta) + 2c - \omega c e^{i\psi} - b\omega e^{i\psi}(e^{-i\theta} - 1)}$$

Scheme III $(b > 0)$

$$\lambda = -\frac{2(\omega - 1)(a + b)(1 - \cos\theta) + 2c(\omega - 1) - \omega c e^{i\psi} - \frac{d}{N} i\omega \sin\theta - b\omega e^{i\psi}(1 - e^{-i\theta})}{2(a + b)(1 - \cos\theta) + 2c - \omega c e^{i\psi} + b\omega e^{-i\psi}(e^{i\theta} - 1)}$$

Scheme IV $(b > 0)$

$$\lambda = -\frac{2(\omega - 1)a(1 - \cos\theta) + 2(c - b)(\omega - 1) - \omega c e^{i\psi} - \frac{d}{N} i\omega \sin\theta + 2b\omega \cos\theta - b\omega e^{i\psi}(e^{i\theta} - 1)}{2a(1 - \cos\theta) + 2(c - b) - \omega c e^{-i\psi} - b\omega e^{-i\psi}(e^{-i\theta} - 1)}$$

Scheme V $(b > 0)$

$$\lambda = -\frac{2(\omega - 1)a(1 - \cos\theta) + 2c(\omega - 1) - \omega c e^{i\psi} - \frac{d}{N} i\omega \sin\theta - 2b\omega \cos\theta - b\omega e^{i\psi}(1 - e^{-i\theta})}{2a(1 - \cos\theta) + 2(c + b) - \omega c e^{-i\psi} + b\omega e^{-i\psi}(e^{i\theta} - 1)}$$

Scheme VI $(b > 0)$

$$\lambda = -\frac{2(\omega - 1)a(1 - \cos\theta) + 2c(\omega - 1) - \omega c e^{i\psi} - \frac{d}{N} i\omega \sin\theta - (\omega - 1)b(1 - e^{i\theta}) - b\omega(e^{i\theta} - 1)}{2a(1 - \cos\theta) + 2c - \omega c e^{-i\psi} - b(1 - e^{-i\theta}) + b\omega e^{-i\psi}(1 - e^{-i\theta})}.$$

Case (i) a and c of same order of magnitude

The stability criterion $|\lambda| \leq 1$ gives

$$\frac{d^2}{N^2} < \frac{4(2-\omega)^2(ac-b^2)}{\omega(4-\omega)} \qquad \text{Schemes I, II and III}$$

$$\frac{d^2}{N^2} < \frac{4[c(2-\omega)-2b]^2(ac-b^2)}{\omega c[c\overline{4-\omega}-4b]} \quad \text{and} \quad [c\overline{2-\omega}-2b] \geq 0$$

Scheme IV $(b \geq 0)$

$$\frac{d^2}{N^2} < \frac{4[c(2-\omega)+2b]^2(ac-b^2)}{\omega c[c(4-\omega)+4b]} \qquad b \geq 0 \qquad \text{Scheme V}$$

$$\frac{d^2}{N^2} - \frac{4db(2-\omega)}{N(4-\omega)} < \frac{4(2-\omega)^2(ac-b^2)}{\omega(4-\omega)} \qquad b \geq 0 \qquad \text{Scheme VI .}$$

The lack of symmetry in scheme VI seems to arise from the lack of symmetry in the representation of the cross-derivative, for example through using an old value for $\phi_{i+1,j}$ while using the current value of $\phi_{i-1,j}$ in equation (3-1a). Scheme V is less restrictive than Schemes I, II and III, but Scheme IV is more restrictive than all these schemes.

All the schemes were tested numerically. ϕ was initially set to 100.0 everywhere, and the scheme was said to have converged at the nth iteration, where n is the smallest integer satisfying

$$\max_k |\phi_k^n - \phi_k^{n-1}| < 10^{-7}$$

where ϕ^n is the vector giving the estimates of ϕ after the nth iteration. The results obtained when $a = c = 1$ are shown in Table 1 .

For the case $b = 0.9$ and $d = 0$ there seems to be little to choose between Schemes I, II and III, but the other schemes are noticeably worse. For Scheme IV the condition $c(2-\omega) > 2b$ is violated and so the iterative scheme is divergent. When b is 0.45 it is predicted that Scheme IV will be stable if and only if $\omega \leq 1.1$. In fact when N is 10 and $\omega = 1.11$ the scheme is convergent, although when N is increased to 50, with ω still 1.11 the scheme is divergent. This illustrates the point that schemes which may appear satisfactory for large step lengths become less and less satisfactory as the step length decreases.

Attention was then turned to d non-zero. The above analysis suggests that when $b = 0.9$, $a = c = 1$ and $\omega = 1.5$

$$|d/N| \leq 0.225 \qquad \text{Scheme I}$$

$$|d/N| \leq 0.439 \qquad \text{Scheme V}$$

for stability. Although, as expected, Scheme V permits larger values of $|d/N|$ than Scheme I the results given in the table are clearly in need of some explanation!

To try to do so the case $a = c = \omega = 1$ and $b = 0$, with $N = 50$ was investigated. The above analysis suggests

$$|d/N| \leq 1.15$$

is a necessary condition for stability in all the senses of section 1. If, however, stability is defined as occurring if all errors eventually decay, then stability will occur if

$$\max_{s,t} |\gamma| < 1$$

where γ satisfies

$$2\gamma - \sqrt{\gamma^2 - \frac{d^2}{4N^2}} \cos\theta - \sqrt{\gamma}\cos\phi = 0$$

with $\theta = \pi s/N$ and $\phi = \pi t/N$, s and t being positive integers satisfying $1 \leq s,t \leq N-1$. As $N \to \infty$ this gives

$$|d/N| < 1.82 .$$

For d/N just greater than 1.15 the iterative scheme works well, but as d/N approaches 1.82 the initial rise in $\max_k |\phi_k^n - \phi_k^{n-1}|$ is so large that it becomes unacceptable, and also causes a large increase in the number of iterations required. Just where the iterative scheme becomes unacceptable is not easy to define, but it is clearly here being ultra-cautious to require $|d/N| < 1.15$, as the intermediate results are not of interest, although allowing it to get too near 1.82 is unacceptable.

Case (ii) An example when a and c are not of the same order of magnitude

It will be assumed that $a\theta^2 \gg c$ ($\Rightarrow a\theta \gg b$ because $b^2 < ac$)

$$b \leq (d/N)$$

$$c \leq (d\theta/N)$$

and that there is an upper limit on the size of N. This situation can arise in the solution of the full potential equations of fluid motion round very highly swept wings using a non-orthogonal grid, in some regions of which the angles between coordinate lines of different families are small.

A von Neumann-type analysis gives approximately,

$$|d| \leq 2aN \tan\frac{\pi}{N}\sqrt{\frac{2-\omega}{\omega}} \qquad \text{Schemes I, II, III and VI} \qquad (3\text{-}4)$$

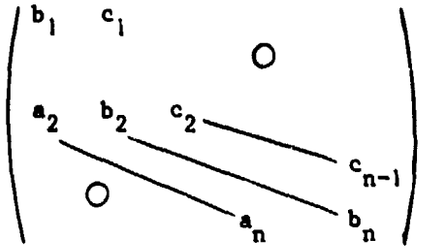
$$|d| \leq 2aN \tan\frac{\pi}{N}\sqrt{\frac{2-\omega}{\omega}}\left[1 + \frac{|b|}{a\tan^2\frac{\pi}{N}(2-\omega)}\right]^{\frac{1}{2}} \qquad \text{Scheme V} \qquad (3\text{-}5)$$

for stability, in all the senses of section 1. No stability criterion is possible for Scheme IV for the following reason. The SLOR method requires the solution of equations

of the form

$$Tx = k \tag{3-6}$$

where T is the $(N - 1) \times (N - 1)$ matrix of form



with $a_i = c_i = -a$ and $b_i = 2(1 - |b| + c)$.
This is done by setting

$$\left.\begin{aligned} \omega_1 &= \frac{c_1}{b_1}, \qquad \omega_i = \frac{c_i}{b_i - a_i\omega_{i-1}} \\ g_1 &= \frac{k_1}{b_1}, \qquad g_i = \frac{k_i - a_i g_{i-1}}{b_i - a_i\omega_{i-1}} \end{aligned}\right\} \tag{3-7}$$

the solution being obtained from

$$x_{N-1} = g_{N-1}, \qquad x_i = g_i - \omega_i x_{i+1}. \tag{3-8}$$

As $a > |b| > c > 0$, there is nothing to prevent $|\omega_i|$ becoming very large occasionally. If $|\omega_i|$ is very large the effect of rounding errors will not be negligible and the scheme will be unstable. A sufficient condition to prevent any SLOR scheme being unstable in this way is to require the matrix, T, to be diagonally dominant, for this ensures $|\omega_i| < 1$ for all i.

The results obtained when $a = 10^4$, $b = 90$ and $c = 1$ are shown in Table 2. Scheme IV appears satisfactory when $N = 10$ and $d = 0$, but was found to be divergent when $N = 50$ and $d = 0$. There is little to choose between Schemes I, II, III and VI but Scheme V is definitely slightly slower (though not so markedly as in the case $a = c = 1$, $b = 0.9$ and $d = 0$).

The symmetrical nature of Schemes I, II and III and the unsymmetrical nature of Scheme VI, with respect to the sign of d, are illustrated, $|d|$ arbitrarily being chosen as 5.5×10^4. The criteria (3-4) and (3-5) suggest that, for stability, when $|d| = 5.5 \times 10^4$, in all the senses of section 1,

$$\omega \leqslant 1.13 \qquad \text{Schemes I, II, III and VI}$$

$$\omega \lesssim \frac{2 + \dfrac{b}{a \tan^2 \frac{\pi}{N}}}{1 + \dfrac{d^2}{4a^2N^2 \tan^2 \frac{\pi}{N}}} \qquad \text{Scheme V} \quad .$$

This implies that all values of ω are permissible with Scheme V - the numerical results support this. However, the restriction on the relaxation factor, ω, does not seem to be quite correct for the other schemes.

To try to explain this, it should first be observed that the criterion ensuring that all errors eventually decay is also given, approximately, by equation (3-4) for Schemes I, II, III and VI. This suggests that as the relaxation factor, ω, approaches its theoretical upper bound the constant, κ, of the second stability criterion, equation (2-10), becomes too large. This permits an unacceptably large initial growth of errors, although, for fixed ω, no error can become arbitrarily large. Thus the results when $|d| = 5.5 \times 10^4$ illustrate the possibly unsatisfactory nature of the second stability criterion. Once again, just where the iterative scheme becomes unsatisfactory is difficult to define.

For a larger value of d - a value of 1×10^5 was taken - with Scheme V, the criterion (3-5) implies, approximately

$$\omega \leq 1.21$$

for stability in all the senses of section 1. However, if it is merely required, for stability, that all errors should eventually decay, the stability criterion may be relaxed to, approximately

$$\omega \leq \frac{8a\left[a \sin^2 \frac{\pi}{N} + B\right]}{d^2/N^2}$$

indicating that all values of the relaxation factor, ω, will be permissible. In the numerical work it was found that values of ω somewhat greater than 1.21 were acceptable, but as ω increased the initial rise in

$$\max_k |\phi_k^n - \phi_k^{n-1}|$$

is so large that the scheme becomes unacceptable. Just where the iterative scheme becomes unsatisfactory is difficult to define. However, as the intermediate results are not of interest, it is clearly too cautious to require $\omega \leq 1.21$, although allowing it to get too large is unsatisfactory.

Discussion

The results of this section have illustrated many of the points made in section 2, concerning possible stability criteria. A von Neumann-type stability analysis has been shown to be of considerable use - although it has shortcomings, some of which have been

illustrated. In most of the section the von Neumann criterion has been seen to be unnecessarily stringent, although in one case it was not quite stringent enough.

The results also suggest ways in which mixed partial derivatives should be handled in second-order elliptic partial differential equations when SLOR is being used to solve the finite-difference equations. Concerning the number of iterations required, and regions of stability, there is little to choose between Schemes I, II and III. However, the calculations for Scheme I will take slightly less time than those for Schemes II and III. Schemes V and VI require more iterations than Scheme I, but Scheme V is stable in circumstances where Schemes I, II and III are unstable. The region of stability of Scheme IV is very much less than the region of stability of Schemes I, II and III, so Scheme IV is of relatively little use. The region of stability of Scheme VI is about the same as the regions of stability of Schemes I, II and III, so Scheme VI is worse than Scheme I. Thus it is recommended that Scheme I should be used whenever possible, but if Scheme I is unstable Scheme V should be tried.

3.2 Example 2 - first derivative

In this example the manner in which instabilities can arise from a boundary condition will be considered. A suitable method for handling some of the first derivative terms which arise in a second-order differential equation will also be indicated.

In the solution of a second-order partial dif erential equation there may be regions in which the equation reduces to, essentia.ly,

$$\phi_{xx} = 0 \tag{3-9}$$

with homogeneous boundary conditions. This will arise through either the coefficients of the other derivatives being small, or through derivatives in directions other than the x direction being small - as can happen in the solution of the full potential equations of fluid motion round very highly swept wings when using a non-orthogonal grid. In such a situation, the manner in which the boundary conditions can cause instabilities is illustrated by the following example: solve (3-9) on $[0,1]$ with

$$\phi = 0 \quad \text{at} \quad x = 0, \quad \phi_x = p\phi \quad \text{at} \quad x = 1, \quad p > 1 \ . \tag{3-10}$$

Take the usual central difference representations of ϕ_x and ϕ_{xx}. Let $\Delta x = 1/N$ and let the subscript i refer to the coordinate direction, x. The finite difference equations may be written in the form:

$$A\phi = \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & \circ & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & \circ & & -1 & 2 & -1 \\ & & & & -2 & 2 - \frac{2p}{N} \end{pmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{N-1} \\ \phi_N \end{pmatrix} = 0 \tag{3-11}$$

where ϕ_i is the estimated value of $\phi(i/N)$.

Several schemes will be considered to solve this equation. They are all of the form:

$$\text{find } v^{n+1} \text{ from } Hv^{n+1} = Gv^{n} \qquad (3\text{-}12)$$

where $H - G = A$, and apply over-relaxation.

Scheme I Here the second derivative will be estimated from the values of ϕ currently being calculated. Hence H is the same as A of equation (3-11) and $G = 0$.

Scheme II Even when equation (3-9) is a good approximation to the full equation, the full equation itself may be much more complex. In such circumstances it is quite likely that some of the first derivative term will be evaluated using values of ϕ currently being calculated while the rest of the first derivative will be estimated using values of ϕ from the previous iteration. This is best understood by writing equation (3-9) as

$$\Phi_{xx} + g\Phi_{x} = g\Phi_{x}$$

where $|g/2N|$ is not necessarily small.

Terms on the left-hand side are to be estimated using the values of ϕ currently being calculated, while the term on the right-hand side is to be found using values of ϕ from the previous iteration. H is now

$$\begin{pmatrix} 2 & -1-g/2N & & & \\ -1+g/2N & \circ & & & \\ & -1+g/2N & 2 & -1-g/2N & \\ & & -2 & 2-(1+g/2N)\,2p/N & \end{pmatrix}$$

and G is

$$\begin{pmatrix} 0 & -g/2N & & & \\ g/2N & \circ & & & \\ & g/2N & 0 & -g/2N & \\ & & 0 & -\frac{g}{2N}\frac{2p}{N} & \end{pmatrix}$$

The method of matrix inversion given in equations (3-6), (3-7) and (3-8) will be used to find y^{n+1} in equation (3-12). To ensure that this is stable, $|g| \leq 2N$ -this ensures that, with the possible exception of the last row, H is diagonally dominant.

Scheme III If $|g| > 2N$ Scheme II will not be satisfactory. However, now write equation (3-9) in the form

$$\phi_{xx} + g\phi_x - N(g - 2N)\phi = g\phi_x - N(g - 2N)\phi$$

and evaluate terms on the left-hand side using the values of ϕ currently being calculated, but terms on the right-hand side using the values of ϕ found during the previous iteration. Then

$$H = \begin{pmatrix} g/N & -1 - g_1/2N & & & \\ -1 + g/2N & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & -1 + g/2N & g/N & -1 - g/2N \\ & & & -2 & g/N - (1 + g/2N)\ 2p/N \end{pmatrix}$$

and $$G = \begin{pmatrix} g/N - 2 & -g/2N & & & \\ g/2N & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & g/2N & g/N - 2 & -g/2N \\ & & & 0 & g/N - 2 - g/2N\ 2p/N \end{pmatrix}.$$

If $g > 2N$ the method given in equations (3-6), (3-7) and (3-8) for finding y^{n+1} in equation (3-12) is stable, because H is diagonally dominant when its last row is ignored.

The stability of the three schemes will now be considered. A von-Neumann analysis gives the following equations for the amplification factors, λ :

Scheme I

$$(\mu - 1)^2(\lambda + \omega - 1) = 0 \qquad (3\text{-}13)$$

Scheme II

$$\lambda = 1 - \frac{(\mu - 1)\omega}{(\mu - 1) + \frac{g}{2N}(\mu + 1)} \qquad (3\text{-}14)$$

Scheme III

$$\lambda = 1 - \frac{(\mu - 1)^2 \omega}{(\mu^2 + 1) + \frac{g}{2N}(\mu^2 - 2\mu - 1)} \qquad (3\text{-}15)$$

where $|\mu| = 1$.

In none of these schemes can $|\lambda|$ be greater than unity for any permitted value of μ , so the von-Neumann stability criterion is satisfied.

The boundary condition at $x = 0$ trivially satisfies the Godunov-Ryabenkii criterion. As the boundary condition at $x = 1$ is being imposed implicitly the finite-difference equation modelling this boundary condition is independent of λ . It is

$$\mu^2 - 1 = \frac{2p}{N}\mu \quad .$$

For modes which decay away from the boundary $|\mu| > 1$, hence the only solution it is necessary to consider is

$$\hat{\mu} = \frac{p}{N} + \sqrt{\frac{p^2}{N^2} + 1}$$

which is real and greater than unity. Substituting this value of μ into equation (3-13) it is found that $\lambda = 1 - \omega$ and so $|\lambda| < 1$. Thus the first scheme satisfies the Godunov-Ryabenkii criterion. As $\hat{\mu} > 1$ and $0 \leqslant \omega \leqslant 2$

$$0 < \frac{(\mu - 1)\omega}{(\mu - 1) + \frac{g}{2N}(\mu + 1)} < 2$$

and so, from equation (3-14), Scheme II also satisfies the Godunov-Ryabenkii criterion.

However, in equation (3-15), $|\lambda|$ will exceed unity if

$$2\frac{g}{2N}(\hat{\mu}^2 - 2\hat{\mu} - 1) + 2(\hat{\mu}^2 + 1) < \omega(\hat{\mu} - 1)^2$$

which, for large $g/2N$, gives, approximately $\hat{\mu} < 1 + \sqrt{2}$ which implies that $p/N < 1$.

The problem given in equations (3-9) and (3-10) was solved numerically in the case $N = 50$, with a relaxation factor of 1.6. In Scheme II g was taken to be 100 and in Scheme III 5000. The results are shown in Table 3, convergence being defined as in section 3.1. They are in good agreement with the above theory.

Discussion

The results of this section have illustrated some of the points made in section 2, about possible stability criteria. As the von Neumann criterion totally ignores the boundary conditions it is possible that even if the von Neumann criterion is satisfied

the scheme may still be unstable as a result of the particular boundary conditions used.

The results also suggest ways in which first derivatives should be handled in second order partial differential equations. It seems safest always to use values of the dependent variable from the previous iteration. If this is not done, too much of the first derivative may be evaluated using the current values of ϕ (Scheme II is unsatisfactory if $|g| > 2N$). If this is avoided (as in Scheme III) the boundary conditions may introduce instabilities.

4 CONCLUSIONS

The question of the stability of iterative schemes has been discussed with the aid of numerical examples. It has been shown that there is in general no satisfactory theoretical definition of stability. This necessarily means that in general there can be no satisfactory criterion for stability.

In many schemes linear equations of the form $\underline{\underline{T}}\,\underline{x} = \underline{k}$ must be solved. The first requirement for stability is that the method employed to solve these equations must be stable. The usual method, if T is tridiagonal, of solving these equations, is given in equations (3-6), (3-7) and (3-8). This method is stable if T is diagonally dominant. However, the results of section 3.2 show that T being diagonally dominant is not always necessary.

If a stable method of solving equations of the form $\underline{\underline{T}}\,\underline{x} = \underline{k}$ is used, a possible fairly simple criterion for stability, is that of von Neumann. The examples illustrate that this criterion can be unnecessarily severe for some problems, and not sufficiently severe for others.

The first example, concerned with the representation of a mixed second derivative (section 3.1), shows that if, in practice, there is a lower bound on the size of the step length taken then the criterion may be too severe. However, in this example, it was only the precise borderlines between practical stability and instability that were incorrectly predicted, all relative trends within and between schemes being correctly predicted.

The second example, concerned with a first derivative (section 3.2), shows that instabilities may arise through the boundary conditions, so that the criterion (which ignores the boundary conditions) may not be severe enough. Thus the criterion indicates that stability may be obtained if the boundary conditions are suitable. As it ignores the boundary conditions it cannot tell which boundary conditions will give stability, and which not.

Thus, while the von-Neumann criterion may be of considerable help in choosing a suitable numerical scheme, it must be applied with great care and its shortcomings kept in mind.

The examples of section 3 suggest how mixed and first derivatives should be handled in SLOR (successive line over-relaxation) schemes. If the mixed derivative is required at the point $x = i\Delta x$, $y = j\Delta y$ then the values of the variable at the points

$x = (i \pm 1)\Delta x$, $y = (j \pm 1)\Delta y$ should be used if possible. However, if stability cannot be obtained with this scheme, a more complex one (Scheme V of section 3.1) may give stability. When evaluation of a first derivative is required, care should be taken if some first derivative terms, on the line currently being updated, are evaluated using values of ϕ currently being calculated, while others are evaluated using values of ϕ from the previous iteration. Stability is more likely to be obtained if first derivatives are always estimated using values of ϕ from the previous iterations.

Table 1

Number of iterations required to solve $\phi_{xx} + 2b\phi_{xy} + \phi_{yy} + d\phi_x = 0$

N	b	d	Scheme	Relaxation factor, ω	Number of iterations
10	0.9	0	I	1.45 (optimum)	31
"	"	"	II	1.46 (optimum)	?
"	"	"	III	1.45 (optimum)	30
"	"	"	IV	1.46	D
"	"	"	V	1.54 (optimum)	158
"	"	"	"	1.45	175
"	"	"	VI	$1.45 \leq \omega \leq 1.60$	70-73
"	0.45	"	IV	1.10	157
"	"	"	"	1.11	182
"	"	"	"	1.20	D
50	"	"	"	1.11	"
10	0.9	6.0	I	1.5	218
"	"	7.5	"	"	D
"	"	"	V	"	91
50	0	55.0	irrelevant as b = 0	1.0	130
"	"	60.0		"	145
"	"	70.0		"	203
"	"	80.0		"	363
"	"	85.0		"	617
"	"	90.0		"	2355
"	"	95.0		"	D

Key: D denotes divergence

Table 2

Number of iterations required to solve $10^4\phi_{xx} + 90\phi_{xy} + \phi_{yy} + d\phi_x = 0$

N	d	Scheme	Relaxation factor, ω	Number of iterations
10	0	I	1.00 (optimum)	6
"	"	II	" "	"
"	"	III	" "	"
"	"	IV	"	15
50	"	"	"	D
10	"	V	1.08 (optimum)	10
"	"	VI	1.00	7
50	$+5.5 \times 10^4$	I	1.05	C
"	"	"	1.06	D
"	"	II	1.05	C
"	"	"	1.06	D
"	"	III	"	C
"	"	"	1.07	D
"	"	VI	1.12	C
"	"	"	1.13	D
"	"	V	$1 < \omega < 1.95$	C
"	-5.5×10^4	I	1.05	"
"	"	I	1.06	D
"	"	II	1.05	C
"	"	"	1.06	D
"	"	III	"	C
"	"	"	1.07	D
"	"	VI	0.98	C
"	"	"	0.99	D
"	"	V	$1 < 1 < 1.95$	C
"	1×10^5	"	1.71	"
"	"	"	1.72	D

Key: D denotes divergence

C denotes convergence, but case not run to full convergence

Table 3

Number of iterations required to solve $\phi_{xx} = 0$ with $\phi = 0$ at $x = 0$ and $\phi_x = p\phi$ at $x = 1$

p	Scheme	Number of iterations
$25\sqrt{2}$	I	43
"	II	C
"	III	D
45	II	C
"	III	D
51	II	C
"	III	C

Key: D denotes divergence

C denotes convergence, but case not run to full convergence

REFERENCES

No.	Author	Title, etc
1	R.D. Richtmyer K.W. Morton	Difference methods for initial value problems. Interscience publishers, 2nd edition (1967)
2	G.G. O'Brien M.A. Hyman S. Kaplan	A study of the numerical solution of partial differential equations. *J. Math. Phys.*, 29, 223-257 (1951)
3	A.R. Mitchell	Computational methods in partial differential equations. Wiley (1969)
4	H.J. Wirz J.J. Smolderen (Eds)	Numerical methods in fluid dynamics. McGraw-Hill (1978)
5	G.D. Smith	Numerical solution of partial differential equations. Oxford University Press, 1965
6	J.L. Siemieniuch I. Gladwell	Analysis of explicit difference methods for a diffusion-convection equation. *Int. J. Num. Meths. Eng.*, 12, 899-916 (1978)
7	K.W. Norton	Stability of finite difference approximations to a diffusion-convection equation. *Int. J. Num. Meths. Eng.*, 15, 677-683 (1980)
8	R.S. Varga	Matrix iterative analysis. Prentice-Hall International (1962)

REPORT DOCUMENTATION PAGE

Overall security classification of this page

UNLIMITED

As far as possible this page should contain only unclassified information. If it is necessary to enter classified information, the box above must be marked to indicate the classification, e.g. Restricted, Confidential or Secret.

1. DRIC Reference (to be added by DRIC)	2. Originator's Reference	3. Agency Reference	4. Report Security Classification/Marking
	RAE TR 81081	N/A	UNLIMITED

5. DRIC Code for Originator	6. Originator (Corporate Author) Name and Location
7673000W	Royal Aircraft Establishment, Farnborough, Hants, UK
5a. Sponsoring Agency's Code	**6a. Sponsoring Agency (Contract Authority) Name and Location**
N/A	N/A

7. Title

An examination of stability criteria for iterative numerical schemes used in solving differential equations

7a. (For Translations) Title in Foreign Language

7b. (For Conference Papers) Title, Place and Date of Conference

8. Author 1. Surname, Initials	9a. Author 2	9b. Authors 3, 4	10. Date	Pages	Refs.
Moore, Katharine			July 1981	33	8

11. Contract Number	12. Period	13. Project	14. Other Reference Nos.
N/A	N/A		Aero 3508

15. Distribution statement

(a) Controlled by – Head of Aerodynamics Dept, RAE

(b) Special limitations (if any) –

16. Descriptors (Keywords) (Descriptors marked * are selected from TEST)

Potential theory*. Numerical analysis*. Iteration*.

17. Abstract

The stability of numerical schemes for solving algebraic finite-difference equations resulting from finite-difference approximations to differential equations is discussed. It is suggested that the von Neumann method together with its stability criterion provides a reasonably simple way of determining stability. However, there are limitations in its applicability, some of which are indicated. The method is tested in two examples and an indication is given of how best to treat first- and mixed-derivative terms occurring in differential equations.

FS910/1